Guidelines for Effectively Assessing Language Learners

Kwansei Gakuin University
Humanities Review
Vol. 24, 2019
Nishinomiya, Japan

Guidelines for Effectively Assessing Language Learners

Shalvin SINGH*

Abstract

This paper outlines a set of guidelines for effectively assessing the proficiency of language learners. These guidelines are intended to address some of the theoretical and practical issues relevant to designing and implementing assessment instruments, particularly for instructors assessing learners in classroom contexts. While learning contexts inevitably vary, the guidelines outlined below are intended to be of interest and relevance to all instructors involved in the education of language learners, and to highlight some of the central issues and justifications for assessing students.

Select an Appropriate Theoretical Model

Theories of language learning underlie all assessment instruments, whether explicitly acknowledged by test designers or not. As such, those who create assessment instruments should first consider and refer to appropriate theoretical models, ways of understanding what language is and what language proficiency looks like, to guide selection of the constructs chosen for assessment. Various theoretical models of language exist, such as Canale and Swain's (1980), Bachman and Palmer's (1996), Celce-Murcia, Dörnyei, and Thurrell's (1995), and Littlewood's (2011). As they each differ substantially in regards to the specific constructs they view as important, as well as the manner in which they conceptualize and interpret communicative competence, selection of a theoretical model is a fundamentally important decision that underpins the basic composition of assessment instruments. Such decisions are crucial as they not only affect the choice

* Instructor of English as a Foreign Language, School of Science and Technology, Kwansei Gakuin University

of constructs selected for assessment but, equally importantly, those that are not selected or viewed as especially relevant.

An example of the manner in which such decisions impact test design would be the importance negotiation of meaning is assigned in L2 oral proficiency assessment instruments, the degree to which this construct is viewed by test designers as a central component of communicative competence. Those who view the construct as essential in assessing oral profiency are likely to use assessment instruments where opportunities for the negotiation of meaning are likely to arise, and favor assessment tasks where there is direct interaction between speakers (Egyud & Glover, 2001). Those, conversely, who believe this construct is not especially important are more likely to view other speaking tasks, such as speeches, structured interviews, and oral readings, as entirely appropriate for assessing oral proficiency. Similarly, a construct such as pronunciation can be understood from numerous, distinct perspectives. Test designers viewing the construct as particularly important are likely to design tasks specifically targeting various aspects of L2 pronunciation, noting the distinction between suprasegmental features (prominence, intonation, connected speech) and segmental features (individual phonemes, lax and open vowels). Various tasks might even be used to effectively assess the construct and an analytical scoring rubric to grade each trait separately would more likely be viewed as appropriate. Others test designers might view pronunciation in a more general way, preferring to assess the construct with reference to concepts such as clarity, comprehensibility, or employ a native-speaker yardstick, and see simpler, holistic rubrics in which pronunciation is conceptualized less thoroughly as useful. At base, theoretical models constitute an argument justifying assessment decisions that favor the assessment of some constructs over others, and the degree and depth with which constructs ought to be analyzed. As such, test designers should be aware that theoretical understandings of language are central to assessment, and form the basis upon which broad concepts (e.g., language ability, writing skills, vocabulary knowledge) are transformed into more manageable components.

Design (Numerous) Assessment Tasks that are Valid

Assessment instruments produce scores and judgments which learners, instructors, institutions and other stakeholders use to evaluate the language proficiency of test takers. The validity claim made by assessment instruments, the extent to which scores and data obtained from the administration of tests can be used to draw conclusions about the ability of learners is, therefore, fundamental to justifying the use of any assessment instrument. As such, it is imperative that validity remain the core concern of test designers, and the central manner of judging

the extent to which an assessment instrument is, or is not, useful.

Test designers often use a variety of terms to judge the quality of assessment instruments, such as authenticity, external validity and directness. However, Messick (1994) argues that it is validity that underlies effective test design and that validity can, and should, be judged primarily by the degree to which an assessment instrument effectively measures a construct. The concepts of construct underrepresentation (failing to measure all variables that constitute a construct) and construct-irrelevant variance (the inclusion of extraneous variables unrelated to a construct) represent central issues in the design of assessment instruments. Test designers should undoubtedly avoid including irrelevant tasks in assessment instruments, as their inclusion prevents tests from effectively measuring learners' language proficiency. An L2 writing test that, for instance, required an understanding of ancient Greek philosophy could face legitimate accusations of construct-irrelevant variance, as such knowledge is not central to demonstrating proficiency in writing. Scores of those better versed in Plato and Aristotle are likely to be higher than those without such knowledge, whether or not their overall L2 writing skills are superior. However, failing to include all relevant tasks necessary for demonstrating language ability is also a matter of concern, and can equally jeopardize the validity of test scores. Assessment instruments often use too narrow a range of tasks to evaluate a construct. Using the TOEIC Listening and Reading test, for example, to evaluate the overall English ability of language learners is a fairly common issue, but similar problems can arise in classroom assessment instruments as well. A test of discourse knowledge specifically targeting one's ability to communicate appropriately in formal situations in which learners were only required to answer multiple choice questions regarding the politeness of speech could understandably be viewed as an example of construct underrepresentation, as most educators would argue that some productive tasks should be included to effectively evaluate the construct. Oftentimes, assessment instruments face simultaneous accusations of both construct underrepresentation and construct-irrelevant variance. A test to measure L2 speaking skills that consisted of a set of ten questions about a learners' political views, could face accusations of both construct underrepresentation and construct-irrelevant variance; the former for only discussing a single topic in a rigid, interview format, the latter for requiring knowledge of politics. To increase validity, test designers should seek to examine the various, relevant aspects of a construct and, in many cases, require that *numerous* assessment tasks be administered, to more successfully evaluate a construct. To give an example, assessment instruments seeking to measure learners' ability to successfully participate in American business meetings could include tasks such as: (a) a role-playing activity in which learners discussed cost-cutting measures, (b) a multiple

choice test examining knowledge of relevant phrases and cultural norms, (c) a 5-minute presentation in which learners deliver advertising pitches, and (d) a listening/writing task in which learners take the minutes of a business meeting. Rather than striving to design single assessment tasks that attempt to examine all possible facets of a construct, by utilizing numerous tasks, single constructs can be examined from various angles, reducing the need for individual tasks to be flawlessly designed and increasing the overall validity of assessment instruments.

Design Assessment Tasks that Promote Positive Washback

Washback – the manner in which testing affects teaching and learning – has consequences both for learners and instructors, prompting changes in behavior, inside and outside the classroom, that impact language learning in significant and lasting ways (Fulcher and Davidson, 2007). While its exact effects are debatable, washback is neither inherently negative nor entirely positive. There are a multitude of ways in which washback influences language learning, potentially affecting motivation, anxiety levels, the time and attention devoted to specific language skills, and the extent to which meaningful, lasting learning takes place (Alderson and Wall, 1993). However, the exact influence of washback undoubtedly varies depending on the specific teaching context in question and the manner in which assessment instruments are designed and used.

Instructors ought to take note of the ways in which assessment tasks focus students on particular aspects of learning, particularly, though not exclusively, in the case of those learners who are not intrinsically motivated (Dörnyei, 1994). Most learners inevitably devote additional time and attention to those aspects of a course of study directly connected to assessment. Doing so is entirely rational and hardly surprising. Given that assessment scores can have consequences that extend far beyond a period of study, learners naturally wish to maximize their scores. Most learners also see higher scores as indicative of learning and intelligence, rarely fretting over a test in which they did especially well. As such, instructors wishing to effectively capitalize upon the realities of washback should attempt to design tasks that assess relevant constructs, to prompt learners to direct attention towards the most important aspects of learning. Failing to do so can have significant consequences for learners. If, for instance, an instructor focuses lessons on developing speaking proficiency, but requires only grammatical knowledge be demonstrated in tests, learners are less likely to view speaking tasks as meaningful or consequential in class. Indeed, such a course's entire design could be criticized as fundamentally flawed. Conversely, were learners informed in advance that assessment instruments will specifically measure speaking ability, and that

assessment tasks will be similar in design to classroom activities, learners are apt to exert greater effort and assign increased importance to such in-class speaking activities. Washback, nevertheless, need not be addressed without consideration of the preferences of the learner. As Rust (2002) notes, assessment tasks which allow students some choice are more likely to be interesting, motivating, and relevant to learners than those over which they have no control. Allowing learners to select topics and areas of focus connected to their interests or learning goals is a simple, but effective way to ensure that assessment can potentially increase learner motivation. Similarly, allowing learners to determine the specific grade value of assessment tasks, whether a speaking or presentation task is worth a greater percentage of their overall grade, would allow learners to devote increased attention to the constructs they view as most relevant.

Avoid Confusing Norm-Referenced Assessment with Criterion-Referenced Assessment

One common error made by many instructors is to assume that an effective classroom assessment instrument has a *good spread* of scores, a few students obtaining very high or low scores, the majority clustered around the middle. This belief represents a fundamental misunderstanding of the purposes of assessment and the specific aims of assessment instruments in distinct contexts. Criterion-referenced assessment and norm-referenced assessment are distinct in terms of their aims, design, and the range of scores one should typically expect, and the two should not be confused (Fulcher and Davidson, 2007). The former is formative, a means of providing feedback for learners and instructors, guides future learning, and is specific to a particular context (such as a classroom context). Conversely, norm-referenced assessment aims to measure learners' general and overall L2 proficiency, typically provides little in the way of meaningful feedback, and measures learners' performance against a specific norm (typically, the performance of a fluent native speaker). Standardized proficiency tests, such as the TOEIC, TOEFL, and IELTS, are common examples of norm-referenced assessment instruments. As classroom assessment is typically criterion-referenced, focused upon measuring the amount of learning that has taken place over the course of a period of study, and aims to provide learners with feedback to guide future learning, a normal distribution of scores should not be expected nor imposed upon learners by instructors. Doing so is akin to punishing learners for achieving a course's goals, i.e., learning that which was specifically taught in a language course. Positively skewed distributions of scores are, therefore, not necessarily an indication of poorly designed assessment instruments, and may rather merely indicate that learners are successfully acquiring

language. Norm-referenced tests, in contrast, typically assess more broadly defined constructs that transcend what might be taught in a specific language course, such as writing skills or speaking ability. Such tests seek to compare the proficiency of learners in relation to a larger population, such as all second language learners, or to native language speakers. As classroom assessment typically centers around evaluating the extent to which learners have acquired the target language introduced in a course of study, its aims are much narrower, and the constructs typically more specific. This suggests that norm-referenced tests ought to play a limited role in typical classroom environments, and that classroom assessment instruments should instead vary according to the specific classroom context, the language targeted for evaluation, the motivation and L2 proficiency of students, and the distinct needs of learners.

Prototype Assessment Instruments Where Possible

Typically, prototyping – the process of analyzing, evaluating, and revising assessment instruments before they are more widely field-tested and implemented – is divided into alpha testing (in-house testing with experts) and beta testing (external testing with small numbers of participants). It is most commonly used to examine the validity of assessment instruments in standardized testing settings, and is combined with statistical analysis to evaluate whether tests need revision before being used with a larger population. However, while such exhaustive processes might be challenging in most classroom contexts, prototyping can still be used in a limited manner to critically examine an assessment instrument prior to its use with a class of language learners (Fulcher and Davidson, 2007).

More practical approaches to prototyping classroom assessment instruments need not be overly onerous or excessively time-consuming. Alpha-testing could simply consist of asking colleagues for advice and feedback regarding assessment tasks, and making use of such feedback to improve the design of instruments. Similarly, having colleagues try out assessment instruments, to check for unclear wording, poorly designed items, overly complex tasks, and the like, is another, realistic manner of improving the quality of assessment instruments and increasing the likelihood they effectively evaluate specific constructs. Beta-testing as well need not be the costly and complex endeavor it is in standardized testing contexts. It could very well simply involve small groups of motivated learners, former students, volunteers, or non-native speaking colleagues of the instructor trying out tests and assessment tasks and noting any problems or issues they encountered. A long-term form of beta-testing could also involve using assessment tasks with an instructor's current set of students, and making revisions based on an analysis of learners'

performance, to improve the quality of assessment instruments for future groups of language learners. In the case of such beta-testing approaches, following the administration of assessment instruments, Rasch and other types of statistical analysis are an effective means of examining the fit of test items, confirming which items are difficult, which are easy, as well as to validate whether there is a good fit between test-takers' ability and the difficulty of items (Beglar, 2010). Although prototyping inevitably requires effort and planning, it is an effective means of ensuring that assessment instruments are evaluated and improved prior to being used to measure learners' language proficiency, as well as a means of improving the quality of previously administered instruments over time.

Conclusion

The guidelines described in this paper represent a brief overview of concepts central to language assessment. Assessment can have a powerful impact upon learners, affecting education and study abroad opportunities, job and career advancement prospects, immigration applications, and a host of other factors central to people's lives. However, its impact in the classroom, while subtler, remains substantial, affecting what learners view as their own unique L2 strengths and weaknesses, the specific skills to which their attention is devoted, and their future language learning goals. It is the intent of these guidelines to assist language educators in designing assessment instruments that effectively meet the needs of language learners and to ensure that such instruments are utilized in a manner that can positively impact students, promote improved learning, foster feedback that is useful and meaningful, and assist learners in achieving their future language learning aims.

References

Alderson, J., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14,* 115-129.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27,* 101-118.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1-47.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6* (2), 5-35

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern*

Language Journal, 78(3), 273-284.

Egyud, G., & Glover, P. (2001). Oral testing in pairs-A secondary school perspective. *ELT Journal 55,* 70-76.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Abingdon, UK: Routledge.

Littlewood, W. (2011). Communicative language teaching: An expanding concept for a changing world. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning,* Volume II (pp.541-547). New York: Routledge.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Rust, C. (2002). The impact of assessment on learning: How can research literature practically help to inform the development of departmental assessment strategies and learner-centered assessment practices? *Active Learning in Higher Education, 3,* 145-158.